명제화된 어트리뷰트 택소노미를 이용하는 나이브 베이스 학습 알고리즘

강대기*, 차경환*

Propositionalized Attribute Taxonomy Guided Naive Bayes Learning Algorithm

Dae-Ki Kang*, Kyung-Hwan Cha*

본 연구는 동서대학교와 동서대학교 지역 혁신 센터(Regional Innovation Center)의 연구비에 의해 지원되었습니다.

요 약

본 논문에서는 명제화된 어트리뷰트 택소노미를 이용하여 간결하고 강건한 분류기를 생성하는 문제를 고려한다. 이 문제를 해결하기 위해 명제화된 어트리뷰트 택소노미(Propositionalized Attribute Taxonomy)를 이용하는 나이브 베이스 학습 알고리즘(Naive Bayes Learner)인 PAT-NBL을 소개한다. PAT-NBL은 명제화된 어트리뷰트들의 택소노미를 선험 지식으로 이용하여 간결하고 정확한 분류기를 귀납적으로 학습하는 알고리즘이다. PAT-NBL은 주어진 택소노미에서 지역적으로 최적의 컷(cut)을 찾아내기 위해 하향식 탐색과 상향식 탐색을 사용한다. 찾아낸 최적의 컷은 명제화된 어트리뷰트 택소노미와 데이터로부터 그에 상응하는 인스턴스 공간(instance space)을 구성할 수 있게 해준다. University of California-Irvine (UCI) 저장소의 기계학습 벤치마크 데이터에 대한 실험 결과를 보면, 제안된 알고리즘이 표준적인 나이브 베이스 학습 알고리즘에 의해 만들어진 분류기들과 비교해 볼 때, 가끔은 보다 간결하고 더 정확한 분류기를 생성해 낸다는 사실을 알 수 있었다.

ABSTRACT

In this paper, we consider the problem of exploiting a taxonomy of propositionalized attributes in order to generate compact and robust classifiers. We introduce Propositionalized Attribute Taxonomy guided Naive Bayes Learner (PAT-NBL), an inductive learning algorithm that exploits a taxonomy of propositionalized attributes as prior knowledge to generate compact and accurate classifiers. PAT-NBL uses top-down and bottom-up search to find a locally optimal cut that corresponds to the instance space from propositionalized attribute taxonomy and data. Our experimental results on University of California-Irvine (UCI) repository data sets show that the proposed algorithm can generate a classifier that is sometimes comparably compact and accurate to those produced by standard Naive Bayes learners.

키워드

명제화, 택소노미, 나이브 베이스 분류기

* 동서대학교 컴퓨터 정보 공학부

접수일자 2008. 09. 09

I. 서 론

기계 학습에서 중요한 목적 중 하나는 이해하기 쉬우면서도 정확하고 간결하며 강건한 분류기를 생성해내는 것이다[1]. 일반적인 기계 학습 과정을 보면, 각각의 인스턴스(instance)는 단순히 어트리뷰트 값들의 튜플(tuple)로 구성된다. 또한 기계 학습 과정에서 생성된 후보 분류기들의 평가와 선택에서는 오캄의 면도날의 원리가 많이 쓰이는 데, 오캄의 면도날의 원리에 따르면, 비슷한 성능을 가지는 두 개의 분류기 중 상대적으로 더 간결한 분류기가 더 선호된다[1].

기계 학습에서 간결한 분류기를 내기 위한 그동안의 연구[2-4]들을 보면, 어트리뷰트 값들 중 서로 비슷한 값들을 하나로 모아서 유사도를 나타낼 수 있고, 해당 분야의 전문가의 지식을 코딩할 수 있는 데, 이러한 결과물이 택소노미를 구성한다. 이러한 과거의 연구들을 보면, 추상적인 값들과 구체적인 값들의 관계를 도시한 택소노미가 더 쉽고 더 이해하기 쉬운 개념(concept)을 익히는 데 도움이 되므로, 위와 같은 목적이 적합하다는 사실을 알 수 있다. 기계 학습에서 택소노미를 응용하는 과거의 연구들을 살펴보면, 크게 두가지로 나뉜다. 하나는 여러 개의 어트리뷰트 값들을 그룹으로 모아서 더 추상적인 값을 만듬으로써 어트리뷰트 값의 택소노미를 만들고 이를 응용하는 연구[4]이며, 다른 하나는 텍스트 문서나 아미노산 시퀀스같은 데이터 집합에서 단어들이나 알파벳 문자들을 모아서 추상적인 값들로 표현하는 연구[5]이다. 이러한 연구들은, 하나의 어트리뷰트(attribute)를 구성하는 여러 개의 어트리뷰트 값(attribute value)들에 대해 그 어트리뷰트 내에서 추상화를 수행하는 연구들이었다.

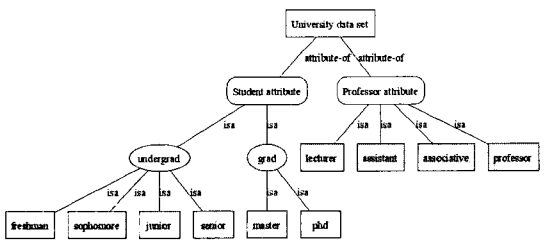

그림 1. 대학교(university) 데이터 집합의 어트리뷰트 값의 택소노미 (명제화가 되지 않고, 추상화되었음)
Fig. 1 Attribute value taxonomy of university data set (no propositionalization applied)

이해를 돕기 위해, 예를 들어, 그림 1에 도시된 택소노미를 가지는 대학교(university)라는 데이터 집합이 있다고 하자. 이 택소노미는 기존의 연구에서처럼 어트리뷰트 내에서 어트리뷰트 값들이 추상화된 것이다. 이 데이터 집합에는 학생(student)라는 어트리뷰트가 있고, 이 어트리뷰트는 1학년(freshman), 2학년(sophomore), 3학년(junior), 4학년(senior), 석사 과정(master), 그리고 박사 과정(phd)이라는 어트리뷰트 값을 가질 수 있다. 또한 교수(professor)라는 어트리뷰트가 있고, 이 어트리뷰트는 강사(lecturer), 조교수(assistant), 부교수(associative), 교수(professor)라는 값들을 가진다고 가정하자. 그림 1의 택소노미에서 각각의 어트리뷰트 값들은 명제화가 되지 않고 어트리뷰트 내에서의 추상화를 통해, 예를 들면 학생 어트리뷰트의 각각의 어트리뷰트 값들은 다시 학부생(undergrad)와 대학원생(grad)이라는 추상화된 값들로 파티션(partition)됨을 알 수 있다.

정리하면, 하나의 데이터 집합은 여러 개의 어트리뷰트들로 구성되어 있고 각각의 어트리뷰트들은 수치값이거나 몇 개의 유한한 어트리뷰트 값을 가지는 데, 과거의 연구들을 보면, 예를 들면 학생 어트리뷰트 값들에서 학부생이나 대학원생 어트리뷰트 값을 도출하는 식으로, 하나의 어트리뷰트 내의 어트리뷰트 값들에 대해 추상화만을 수행하고, 서로 다른 어트리뷰트들이 가지는 어트리뷰트 값들에 대해서는 추상화를 고려하지 않았다. 그러나, 현실 세계에서는 서로 다른 어트리뷰트들로부터 유래된 어트리뷰트 값들 간에도 유사성이 존재하며, 이를 추상화한 값이 주어진 문제에 더 적합한 경우들이 많이 존재한다. 예를 들어, 위의 대학교 데이터 집합에서 특정 분야의 학술 논문을 리뷰할 수 있는 사람들에 대한 추론(교수 집단과 대학원생 집단)이나, 대학교에 입학했거나 직장으로 다닌지 2년이 안되는 사람들에 대한 추론(1학년, 2학년, 강사 및 조교수 중 일부)을 하는 경우에는, 서로 다른 어트리뷰트 내의 어트리뷰트 값들의 부분집합들을 합하여 추상화한 값이 더 적합하다.

본 논문에서는, 이러한 문제를 해결하기 위한 방법 중 하나로서, 하나의 어트리뷰트 값을 하나의 어트리뷰트로 명제화(propositionalization)하고 이 명제화된 어트리뷰트들을 기반으로 명제화된 어트리뷰트 택소노미(Propositionalized Attribute Taxonomy; PAT)를 만들어 기

계 학습 분야에 대해 적용해보고자 한다. 그림 2는 학술 논문을 리뷰할 수 있는 사람에 대해 적합한 명제화된 어트리뷰트의 택소노미이다.

본 논문에서는, 명제화된 어트리뷰트 택소노미인 PAT를 이용하기 위해, 기존의 나이브 베이스 학습 알고리즘(Naive Bayes Learner)을 확장한 Propositionalized Attribute Taxonomy guided Naive Bayes Learner(PAT-NBL)를 소개한다. PAT-NBL은 나이브 베이스 학습 알고리즘을 택소노미를 이용할 수 있도록 확장한 것으로, 추상화(abstraction)를 통한 상향식 탐색과 정련화(refinement)를 통한 하향식 탐색을 사용하여 주어진 택소노미에서 최적의 컷(cut)를 찾는다. 주어진 데이터 집합들에 대해 항상 택소노미가 존재하지는 않으므로, 이 문제를 해결하기 위해 주어진 데이터에서 기계 학습에 유리한 택소노미를 자동으로 생성하기 위해, 각각의 어트리뷰트들을 클래스 조건 분포(class conditional distribution)에 따라 계층 응집 클러스터링(hierarchical agglomerative clustering; HAC)해주는 알고리즘[5]을 사용하였다.

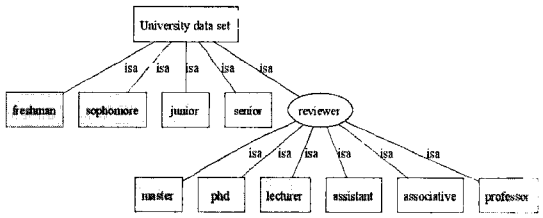

그림 2. 대학교(university) 데이터 집합의 명제화된 어트리뷰트 택소노미 (명제화된 후, 추상화되었음)
Fig. 2 Propositionalized attribute taxonomy of university data set (propositionalized)

PAT-NBL 알고리즘을 평가해 보기 위해, 우리는 University of California-Irvine (UCI) 저장소[6]의 벤치마크 데이터들에 대해 비교 실험을 수행하였다. 실험 결과, PAT-NBL 알고리즘을 통해 생성된 분류기들은 표준적인 나이브 베이스 분류기에 의해 만들어진 분류기들과 비교해 볼 때, 가끔은 보다 간결하고 더 정확하다는 사실을 알 수 있었다.

II. 명제화된 어트리뷰트 택소노미 (Propositionalized Attribute Taxonomy)

본 단원에서는 PAT-NBL 알고리즘을 서술하기 전에 기본적인 개념들에 대해 먼저 서술하고자 한다.

$A=\{a_1,a_2,...,a_I\}$과 같이 A를 유한한 어트리뷰트들의 집합으로 정의하고, $V_i=\{v_i^1,v_i^2,...,v_i^{m_i}\}$를 ai 에 속해 있는 서로 다른 어트리뷰트 값들의 유한한 집합으로 정의하자. 여기서, vij는 어트리뷰트 ai의 j 번째 어트리뷰트 값이며 I 은 어트리뷰트의 개수이고 mi 은 ai에 속한 어트리뷰트 값들의 개수이다. $C=\{c_1,c_2,...,c_n\}$는 서로 겹치지 않는 클래스 레이블들의 집합이라고 하자. 인스턴스 I는 어트리뷰트 값들의 고정된 개수로 구성된 튜플로 $I\in V_1\times V_2\times...\times V_I$ 로 나타낼 수 있다. 데이터 집합 D는 인스턴스들과 각각의 인스턴스에 상응하는 클래스 레이블의 집합으로 $D\subseteq V_1\times V_2\times...\times V_I\times C$ 이다.

Definition 1 (명제화). 명제화(propositionalization)는 함수 $f: V_i\rightarrow\tilde{A}$로서, 어트리뷰트 $a_i\in A$ 에 대응하는 각각의 어트리뷰트 값 $v_i^j\in V_i$에 대해 새로운 부울리안 어트리뷰트 $\tilde{a}_i\in\tilde{A}$를 출력하는 것이다.

명제화된 어트리뷰트 $\tilde{a}_i$의 어트리뷰트 값 $\tilde{v}_i^j$은 부울리안 값으로 $\{True,False\}$에 포함되며, 명제화된 데이터 집합 $\tilde{D}$는 $\tilde{D}\subseteq\tilde{V}_1\times\tilde{V}_2\times...\times\tilde{V}_I\times C$로 정의된다.

명제화된 인스턴스 $\tilde{I}$의 어트리뷰트 $\tilde{a}_i$는, 원래의 인스턴스 I 가 상응하는 어트리뷰트 값 v_i^j를 가질 때, True 값을 가진다.

$\tilde{T}$를 A에서 명제화된 부울리안 어트리뷰트 $\tilde{A}$ 위에 정의된 명제화된 어트리뷰트 택소노미(PAT)로 정의하자. 그러면, Root($\tilde{T}$)를 $\tilde{T}$의 루트 노드로 표시할 수 있고, $\tilde{T}$의 리프 (leaf) 노드들을 Leaves($\tilde{T}$)$\subseteq\tilde{A}$로 표시할 수 있다. 그리고, 택소노미의 내부 노드들은 $\tilde{A}$의 추상적인 값들에 대응된다.

예를 들어, 그림 1의 대학교 데이터 집합은 학생과 교수라는 두 개의 어트리뷰트를 가지고, 학생은 { 1학년, 2학년, 3학년, 4학년, 석사과정, 박사과정 } 이라는 여섯 개의 어트리뷰트 값을 가지며, 교수는 {강사, 조교수, 부

교수, 교수 } 이라는 네 개의 어트리뷰트 값들을 가진다. 데이터 집합의 인스턴스들의 일부는 예를 들어 표 1과 같을 것이다.

표 1. 대학교 데이터 집합의 일부
Table. 1 Instances of university data set

student	professor
freshman	lecturer
sophomore	professor
freshman	associative
senior	assistant

이 데이터 집합이 명제화(propositionalization)가 되면, 명제화된 데이터 집합은 10 개의 어트리뷰트를 가지게 된다. 데이터 집합들은 표 2와 같이 명제화된 형태를 취하게 된다.

표 2. 명제화된 대학교 데이터 집합의 일부
Table. 2 Propositionalized instances of university data

1	2	3	4	MS	PhD	전강	조교수	부교수	정교수
T	F	F	F	F	F	T	F	F	F
F	T	F	F	F	F	F	F	F	T
T	F	F	F	F	F	F	F	T	F
F	F	F	T	F	F	F	T	F	F

그림 2는 이렇게 명제화된 어트리뷰트들을 가지는 데이터 집합에 대해 택소노미를 구한 예인 것이다.

Haussler[7]의 정의에 따르면, 택소노미 $\tilde{T}$를 통하는 컷(cut) γ는 다음과 같이 정의된다.

Definition 2 (컷). 컷(cut) γ는 택소노미 $\tilde{T}$ 안에 포함된 노드들의 부분집합으로 다음의 두가지 특성들을 만족한다.

- 임의의 리프 노드 x∈Leaves($\tilde{T}$)에 대해, x는 γ에 포함되거나 x는 γ에 포함된 노드의 자손(descendant)이다.
- γ에 포함된 임의의 두 노드들 x, y에 대해, x는 y의 선조(ancestor)나 자손(descendant)이 아니다.

$\tilde{T}$의 컷 γ는 명제화된 어트리뷰트들 $\tilde{A}$의 분할을 이끌어낸다.

Definition 3 (추상화와 정련화). 컷 γ의 적어도 하나의 노드 v를 그 자손으로 교체해서 컷 $\hat{\gamma}$를 얻었을 경우, 컷 $\hat{\gamma}$은 컷 γ를 정련화(refinement)한 것이다. 반대로, γ는 $\hat{\gamma}$의 추상화(abstraction)이다.

Definition 4 (명제화된 인스턴스 공간). 택소노미 $\tilde{T}$에 대해 하나의 컷 γ를 선택한 경우, 이에 상응하는 명제화된 인스턴스 공간 $\tilde{I}_\gamma$이 유도된다. 만일 컷 γ에 포함된 하나의 노드가 Leaves($\tilde{T}$)에 포함되지 않는다면, 유도된 인스턴스 공간 $\tilde{I}_\gamma$는, 원래의 인스턴스 공간 I 이 명제화되어 생성된 인스턴스 공간 $\tilde{I}$의 추상화이다.

데이터 집합 D, 택소노미 $\tilde{T}$ 그리고 그에 대응되는 컷들을 통해, 우리는 인스턴스 공간에 대한 우리의 정의를 확장하여 명제화된 인스턴스 공간의 서로 다른 레벨의 추상화를 통해 유도된 인스턴스 공간들을 포함시킬 수 있다. 비슷하게, 인스턴스 공간 $\tilde{I}_\gamma$에서 생성된 가설 $\hat{\gamma}$는 인스턴스 공간 $\tilde{I}_\gamma$에 대응되는 가설 γ의 정련화이다. PAT-NBL 알고리즘은 이렇게 유도된 인스턴스 공간에서 수행된다.

Ⅲ. PAT-NBL 알고리즘

명제화된 어트리뷰트 택소노미(PAT)와 명제화된 데이터에서 분류기를 학습하는 문제는 데이터로부터 학습을 하는 문제의 확장이다. 원래의 데이터 집합 D는 클래스 레이블이 붙은 인스턴스 $\langle I, C\rangle$의 집합이다. 분류기(classifier)는 $h : I \rightarrow C$라는 함수 형태의 가설이고, 가설 공간 H는 가설 언어 또는 함수들의 매개변수들로 표현되는 가설들의 집합이다. 이러한 정의들에 따라, 데이터 집합 D로부터 분류기를 학습하는 작업은 주어진 기준을 만족하는 가설 $h \in H$를 유도하는 작업이다.

비슷하게, 명제화된 어트리뷰트 택소노미(PAT)와 명제화된 데이터에서 분류기를 학습하는 문제는 다음과 같이 서술될 수 있다. 주어진 명제화된 어트리뷰트 택소

노미 $\tilde{T}$와 명제화된 데이터 집합 $\tilde{D}$에 대해, 우리의 목표는 분류기 $h_\gamma : \tilde{I}_\gamma \rightarrow C$를 유도하는 것으로, 여기서 γ^*는 주어진 기준을 최대화하는 컷이다.

명제화된 어트리뷰트 택소노미(PAT)에 의해 안내되는 나이브 베이스 학습 알고리즘 (PAT-NBL)은 가설 공간 내에서 상향식 또는 하향식으로 다단계로 구성되어 있는 PAT에 대해 언덕 오르기 탐색을 수행하는 알고리즘이다. 이를 위해 PAT-NBL은 주어진 PAT에 근거하여 카운트를 계산하는 모듈과 이러한 카운트를 근거로 나이브 베이스 학습기를 구성하는 모듈로 구성되어 있다.

PAT-NBL 알고리즘에서 주어진 PAT 상에서 지역적으로 최적인 컷을 탐색하기 위해서는, 특정 컷에서 생성되는 모델을 평가할 수 있는 기준이 있어야 한다. 알고리즘은 주어진 평가 기준을 가지고 탐색을 수행하며 후보 컷들 중 기존의 컷보다 주목할만한 향상을 보이는 컷이 하나도 없을 때까지 탐색을 반복한다.

본 논문에서는 모델 평가에 많이 쓰이는 다음의 세 가지 기준을 가지고 실험을 수행하였다.

- 조건부 로그 우도 (Conditional Log-Likelihood; CLL) [8]
- 조건부 최소 코드 길이 (Conditional Minimum Description Length; CMDL)
- 조건부 아케이케 정보 기준 (Conditional Akaike Information Criteria; CAIC)

v_j가 주어진 데이터 D에 속한 인스턴스 $d_j \in D$의 어트리뷰트 값들의 집합이고, $c_j \in C$가 클래스 레이블의 집합 C의 원소로 d_j에 대한 클래스 레이블이라고 할 때, 주어진 가설 B의 조건부 로그 우도(CLL)는 다음과 같다.

$$CLL(B|D) = |D| \sum^{|D|} \log\{P_B(c|v)\}$$
$$= |D| \sum^{|D|} \log\left\{ \frac{P_B(c) P_B(v|c)}{\sum_{c_i}^{|C|} P_B(c_i) P_B(v|c_i)} \right\}$$

나이브 베이스 분류기의 경우, 위의 값은 다음과 같이 추산할 수 있다.

$$CLL(B|D) = |D| \sum^{|D|} \log\left\{ \frac{P_B(c) \prod^{v_i \in v} P_B(v_i|c)}{\sum_{c_i}^{|C|} P_B(c_i) \prod^{v_i \in v} P_B(v_j|c_i)} \right\}$$

또한 조건부 최소 코드 길이(CMDL) 값은 다음과 같이 구해진다.

$$CMDL(B|D) = - CLL(B|D) + \left\{ \frac{\log(|D|)}{2} \right\} size(B)$$

여기서 size(B)는 가설 B의 크기이다.

조건부 아케이케 정보 기준(CAIC)[9] 값은 단순히 다음과 같다.

$$CAIC(B|D) = - CLL(B|D) + size(B)$$

PAT-NBL:
begin
1. 입력 : 명제화된 데이터 집합 $\tilde{D}$와 명제화된 어트리뷰트 택소노미 $\tilde{T}$
2. 컷 γ를 $\tilde{T}$의 leaf 노드들인 Leaves($\tilde{T}$)로 초기화한다.
3. 클래스 조건부 빈도 카운트를 계산하고 가설 h_γ를 생성한다.
4. 컷 γ에 아무런 변화가 없거나 $|\gamma| < 1$이 되는 경우까지 다음을 반복한다.
   가. $\bar{\gamma} \leftarrow \gamma$ // 현재의 컷을 임시 저장함
   나. $p_\gamma \leftarrow \gamma$ 안에 포함된 모든 노드들의 부모들의 집합 // p는 부모(parent)를 의미함
   다. p_γ에 포함되는 임의의 노드 $v \in p_\gamma$에 대해 다음을 수행한다:
      1) γ 안에 있는 노드들 중 v의 자식 노드들을 v로 바꾸어, γ의 추상화된 컷 γ_v를 생성한다.
      2) 추상화된 컷 γ_v에 상응하는 가설 h_{γ_v}를 구성한다.
      3) 만일 $CMDL(h_{\gamma_v}) > CMDL(h_{\bar{\gamma}})$이면 $\bar{\gamma}$를 γ_v로 변경한다.
   라. 만일 $\bar{\gamma} \neq \gamma$이면 $\gamma \leftarrow \bar{\gamma}$로 설정한다.
5. 출력 : h_γ
end

그림 3. 조건부 최소 코드 길이(CMDL)를 모델 평가 기준으로 사용하고 추상화를 통한 상향식 탐색을 수행하는 PAT-NBL 알고리즘의 의사 코드
Fig. 3 Pseudo code of PAT-NBL algorithm that adapts CMDL and abstraction

표 3. UCI 벤치마크 데이터들에 대한 실험 결과
Table. 3 Experimental results on UCI benchmark data sets

Data	NBL (original)		PAT-NBL(추상화/CLL)		PAT-NBL(추상화/CMDL)		PAT-NBL(추상화/CAIC)	
	정확도	크기	정확도	크기	정확도	크기	정확도	크기
Anneal	96.66±1.18	768	89.87±1.97	54	89.87±1.97	54	89.87±1.97	54
Autos	71.71±6.17	798	66.83±6.45	791	53.17±6.83	231	55.12±6.81	252
Balance-scale	70.72±3.57	27	75.20±3.39	24	75.20±3.39	24	75.20±3.39	24
Breast-cancer	71.68±5.22	104	73.08±5.14	102	72.73±5.16	66	72.73±5.16	66
Breast-w	97.00±1.27	60	97.28±1.21	58	97.28±1.21	58	97.28±1.21	58
Dermatology	97.81±1.50	906	98.09±1.40	900	98.36±1.30	564	98.09±1.40	582
Heart-statlog	83.33±4.45	46	84.07±4.36	44	84.07±4.36	44	84.07±4.36	44
Hepatitis	85.16±5.60	74	84.52±5.70	72	85.16±5.60	54	83.87±5.79	60
Hypothyroid	98.62±0.37	272	97.91±0.46	268	97.91±0.46	268	97.91±0.46	268
Ionosphere	90.60±3.05	292	89.46±3.21	290	92.31±2.79	110	92.02±2.83	112
Kr-vs-kp	87.89±1.13	150	85.01±1.24	148	77.72±1.44	96	81.88±1.34	100
Labor	91.23±7.34	72	92.98±6.63	70	89.47±7.97	48	89.47±7.97	48
Mushroom	95.83±0.43	252	94.25±0.51	250	96.66±0.39	156	94.76±0.48	182
Segment	91.52±1.14	1204	91.04±1.16	1197	88.83±1.28	651	88.83±1.28	658
Sonar	85.58±4.77	164	86.06±4.71	162	83.65±5.03	70	84.13±4.97	72
Splice	95.52±0.72	864	95.64±0.71	861	91.88±0.95	213	51.58±1.73	21
Vehicle	62.65±3.26	296	62.29±3.27	292	59.34±3.31	188	61.35±3.28	200
Vote	90.11±2.80	66	88.51±3.00	64	88.74±2.97	52	88.51±3.00	64
Waveform	80.74±1.09	393	81.24±1.08	390	80.14±1.11	159	80.54±1.10	168
Zoo	93.07±4.95	259	96.04±3.80	252	96.04±3.80	245	96.04±3.80	252

언덕 오르기 탐색으로 주어진 택소노미에서 지역적으로 최적인 컷을 찾기 위해서는 두 가지 방법을 생각할 수 있다. 하나는 상향식 탐색이고, 다른 하나는 하향식 탐색이다. PAT-NBL은 컷을 추상화하는 상향식 탐색과 컷을 정련화하는 하향식 탐색을 둘 다 사용한다. 그림 3은 추상화를 사용하는 PAT-NBL 알고리즘을 도시한 것이다. 컷 γ는 처음에는 명제화된 어트리뷰트 택소노미(PAT)의 leaf 노드들인 Leaves($\tilde{T}$)로 초기화된다. 알고리즘이 진행함에 따라, 컷 안의 노드들은 자신들의 부모 노드로 추상화된다. 알고리즘은 이러한 추상화를 통해 컷이 더 이상 주목할만큼 향상되지 않을 때까지 추상화를 반복한다.

그림 3은 조건부 최소 코드 길이(CMDL)를 모델 평가 기준으로 사용하고 추상화를 통한 상향식 탐색을 수행하는 PAT-NBL 알고리즘의 의사 코드(pseudo code)이다.

Ⅳ. 실험 결과 및 고찰

우리는 University of California-Irvine (UCI) 저장소[6]의 벤치마크 데이터 집합[1]들에 대해 비교 실험을 수행하였다. 데이터의 어트리뷰트 중 숫자와 같은 연속적인 값을 가지는 어트리뷰트에 대해서는 유한한 값들의 집합으로 변환하는 이산화(discretization)를 수행하였다.

네 개의 서로 다른 설정에 대해 비교 실험을 하였는데, 첫번째는 표준적은 나이브 베이스 학습기를 사용한 것(NBL)이며, 두번째는 추상화와 조건부 로그 우도(CLL)를 사용하는 PAT-NBL, 세 번째는 추상화화 조건부 최소 코드 길이(CMDL)를 사용하는 PAT-NBL, 그리고 네번째는 추상화와 조건부 아케이케 정보 기준

1) http://www.cs.waikato.ac.nz/ml/weka/index_datasets.html

(CAIC)을 사용하는 PAT-NBL이다. 개개의 알고리즘의 평가를 위해 10 폴드 교차 검증 방법(10-fold cross-validation)을 사용하였다. 각각의 경우마다 대해 택소노미는 학습 데이터로부터 계층 응집 클러스터링 (hierarchical agglomerative clustering; HAC)해주는 알고리즘[5]으로 생성되어 사용되었다.

표 3는 각각의 실험 설정에 대한 분류기의 정확도와 분류기의 크기를 도시한 것이다. 실험 결과에 따르면, 정확도만을 고려해 보면, 서로 다른 네 개의 설정 중 어느 것도 전반적으로 우월한 결과를 내진 않았다는 것이다. 그러나, 제안된 알고리즘들이 표준적인 나이브 베이스 알고리즘과 비슷한 정확도를 보였으며 가끔은 보다 간결하고 더 정확한 분류기를 생성해 낸다는 사실을 알 수 있었다. 그리고, 분류기의 크기를 고려해 보면, 제안된 알고리즘들이 표준적인 나이브 베이스 알고리즘보다 언제나 더 좋은 결과를 보임을 알 수 있다.

V. 결론

본 논문에서는 명제화된 어트리뷰트 택소노미를 이용하여 간결하고 강건한 분류기를 생성하는 문제를 고려해 보았다. 이를 위해 명제화된 어트리뷰트 택소노미(Propositionalized Attribute Taxonomy)를 이용하는 나이브 베이스 학습 알고리즘(Naive Bayes Learner)인 PAT-NBL을 제시하였다. PAT-NBL은 명제화된 어트리뷰트들의 택소노미를 선험 지식으로 이용하여 간결하고 정확한 분류기를 귀납적으로 학습하는 알고리즘이다. PAT-NBL은 주어진 택소노미에서 지역적으로 최적의 컷(cut)을 찾아내기 위해 하향식 탐색과 상향식 탐색을 사용한다. 찾아낸 최적의 컷은 명제화된 어트리뷰트 택소노미와 데이터로부터 그에 상응하는 인스턴스 공간(instance space)을 구성할 수 있게 해준다. University of California-Irvine (UCI) 저장소의 기계학습 벤치마크 데이터에 대한 실험 결과를 보면, 제안된 알고리즘이 표준적인 나이브 베이스 학습 알고리즘에 의해 만들어진 분류기들과 비교해 볼 때, 가끔은 보다 간결하고 더 정확한 분류기를 생성해 낸다는 사실을 알 수 있었다.

감사의 글 (Acknowledgement)

저자들은 논문을 자세히 심사하고 리뷰해 주신 심사위원님들께 감사드립니다.

참고문헌

[1] M. J. Pazzani, S. Mani, and W. R. Shankle. Beyond concise and colorful: Learning intelligible rules. In Knowledge Discovery and Data Mining, pages 235 - 238, 1997.

[2] J. Han and Y. Fu. Exploration of the power of attribute-oriented induction in data mining. In Advances in Knowledge Discovery and Data Mining. 1996.

[3] M. G. Taylor, K. Stoffel, and J. A. Hendler. Ontology based induction of high level classification rules. In Data Mining and Knowledge Discovery, 1997.

[4] J. Zhang and V. Honavar. Learning decision tree classifiers from attribute value taxonomies and partially specified data. In Proc. of the Twentieth International Conference on Machine Learning, 2003.

[5] D.-K. Kang, J. Zhang, A. Silvescu, and V. Honavar. Multinomial event model based abstraction for sequence and text classification. In Proc. of 6th International Symposium on Abstraction, Reformulation and Approximation, pages 134 - 148, 2005.

[6] C.L. Blake and C.J. Merz. UCI repository of machine learning databases, 1998.

[7] D. Haussler. Quantifying inductive bias: AI learning algorithms and Valiant's learning framework. Artificial intelligence, 36:177 - 221, 1988.

[8] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. Mach. Learn., 29(23):131 - 163, 1997.

[9] H. Akaike. Information theory and an extension of the maximum likelihood principle. In Proceedings of Second International Symposium on Information Theory, pages 267 - 281, 1973.

저자소개

강대기 (Kang, Dae-Ki)

1992년 한양대학교 전자계산학과 졸업
1994년 서강대학교 전자계산학과 (이학 석사)
1994년~1999년 한국전자통신연구원 (연구원)
2006년 Iowa State University (PhD in Computer Science)
2007년 2월~2007년8월 국가보안기술연구소 (선임연구원)
2007년 9월~현재 동서대학교 컴퓨터정보공학부 전임강사
※관심분야 : 기계학습, 관계학습, 통계적그래피컬모델, 온톨로지학습, 침입탐지, 웹방화벽, 웹마이닝, 컴퓨터 비전

차경환(Cha, Kyung-Hwan)

1985년 부경대학교 전자통신공학과 (공학사)
1990년 부경대학교 전자통신공학과 (공학석사)
1996년 부경대학교 전자공학과(공학박사)
1990년~1995년 LG전자 DAC 연구소 주임연구원
1995년~현재 동서대학교 컴퓨터정보공학부 교수
※관심분야 : 임베디드시스템, 디지털신호처리